文章编号:1674-8190(2021)03-085-10

基于深度确定性策略梯度算法的战机规避中距空空导弹研究

宋宏川[1],詹浩[1],夏露[1],李向阳[2],刘艳[1]
(1.西北工业大学 航空学院,西安 710072)
(2.西安地平线电子科技有限公司,西安 710072)

摘 要: 飞机规避中距空空导弹的逃逸机动策略对于提高战斗机的生存力至关重要。针对深度确定性策略梯度算法训练智能体学习飞机规避导弹的逃逸机动策略进行研究。以飞机导弹相对态势参数等作为智能体的输入状态,飞机控制指令作为智能体的输出动作,导弹飞机追逃模型作为智能体的学习环境,设计由相对态势和飞行参数构成的成型奖励以及由交战结果组成的稀疏奖励,实现从状态参数到控制量端到端的逃逸机动策略。通过与四种基于专家先验知识的典型逃逸机动攻击区仿真验证对比,结果表明:智能体实现的逃逸策略攻击区仅次于置尾下降攻击区,该策略对飞机规避导弹先验知识的依存度最低。

关键词: 导弹规避;逃逸机动策略;深度确定性策略梯度;深度强化学习
中图分类号: V212.1; E91; E926.3 **文献标识码:** A
DOI: 10.16615/j.cnki.1674-8190.2021.03.11 **开放科学(资源服务)标识码(OSID):**

The Study on a Fighter Against a Medium-range Air-to-air Missile Based on Deep Deterministic Policy Gradient Algorithm

SONG Hongchuan[1], ZHAN Hao[1], XIA Lu[1], LI Xiangyang[2], LIU Yan[1]
(1. School of Aeronautics, Northwestern Polytechnical University, Xi'an 710072, China)
(2. Skyline Technologies, Xi'an 710072, China)

Abstract: The evasive maneuver strategy for a fighter against a medium-range air-to-air missile is crucial to improving aircraft survivability. In this paper, the deep deterministic policy gradient algorithm to train the agent to learn the evasive maneuver strategy is studied. The missile-aircraft engagement model parameters are the input states. The aircraft control commands are taken as the output actions. The missile-aircraft pursuit-evasion model is taken as the learning environment. The shaping reward, including engagement model parameters and flight parameters, and the sparse reward of the engagement results are designed. Finally, the end-to-end evasive maneuver strategy from the state parameters to the aircraft control variables is realized. The attack zones of four classic evasive maneuvers based on prior knowledge by simulating are compared. It is proved that the evasion strategy developed in this paper is second only to the tail dive maneuver. However, this strategy has the lowest dependence on the specialized domain knowledge of missile evasion.

Key words: missile evasion; evasive maneuver strategy; deep deterministic policy gradient; deep reinforcement learning

收稿日期: 2021-02-08; **修回日期:** 2021-03-15
基金项目: 国家自然科学基金(11672236)
通信作者: 宋宏川, hongchuanbox@163.com
引用格式: 宋宏川, 詹浩, 夏露, 等. 基于深度确定性策略梯度算法的战机规避中距空空导弹研究[J]. 航空工程进展, 2021, 12(3): 85-94.
SONG Hongchuan, ZHAN Hao, XIA Lu, et al. The study on a fighter against a medium-range air-to-air missile based on deep deterministic policy gradient algorithm[J]. Advances in Aeronautical Science and Engineering, 2021, 12(3): 85-94. (in Chinese)

0 引 言

现代空战根据雷达探测范围和使用武器类型可分为超视距空战和近距空战。随着机载雷达和空空导弹性能的提升,空战在超视距阶段结束战斗的比例已从20世纪80年代的不足30%上升到21世纪初的超过50%[1]。因此,如何在有限的机动能力下,使用高效的逃逸机动策略提高战斗机对中距空空导弹的规避、逃逸能力,对于提高其空战生存力至关重要[2]。

飞机规避导弹问题是追逃对策的一种,导弹是追击者,根据导引律策略追击飞机;飞机是逃逸者,决策最优控制策略逃脱导弹的追击。传统的飞机规避导弹通常采用专家系统法[3-7]、微分对策法[8-10]、最优控制法[11-13]以及模型预测法[14-15]求解最优或次优逃逸机动。专家系统法极其依赖人类专家的先验知识,当导弹或飞机子系统变化时,人类专家需要分析新的子系统并再次给出新逃逸机动策略。微分对策、最优控制以及模型预测方法都依赖于明确完备的数学模型,需要对复杂的微分方程求解析或数值解。飞机规避导弹问题包括众多复杂非线性系统,各个子系统建模不免存在误差,这无疑加大了以上方法求解飞机规避导弹策略的难度。

近年来随着人工智能的发展,强化学习和深度神经网络相结合衍生出了一系列无需建模,只通过端到端学习,便能够实现从原始输入到输出的直接控制算法[16]。深度确定性策略梯度算法(Deep Deterministic Policy Gradient,简称DDPG)是其中一种可应用于连续动作空间的免模型算法[17]。国内外已经将该算法应用于不同的领域。WANG M等[18]利用DDPG算法研究了平面小车的追逃问题;S. YOU等[19]和R. Cimurs等[20]利用该算法研究了智能体在避开动态和静止障碍物的同时,追击目标的导航问题。上述研究中动态障碍物的动力学和运动学模型相对简单,相比于逃逸者,追击者并没有速度和机动性的绝对优势且追击者并未采用有效的追击策略。范鑫磊等[21]将DDPG算法应用于导弹规避决策训练,仿真验证了四种典型初始态势下逃逸策略的有效性,但在其研究中,空空导弹和规避飞机均采用简化模型,未考虑导弹导引律、飞机导弹气动模型以及飞机导弹相对运动模型,且其初始态势的范围相对较少,未与典型的逃逸策略进行对比,未能对DDPG算法学习到的逃逸策略有效性作出更精确的评价。

针对以上问题,本文基于DDPG算法,构建一套导弹规避训练系统。首先建立导弹飞机追逃模型,包括飞机导弹质点模型(考虑气动特性和推力特性)、空空导弹的导引律和杀伤率模型以及飞机导弹相对运动模型;再介绍DDPG算法并设计基于DDPG算法的导弹追逃问题奖励;然后将导弹追逃问题建模为基于DDPG算法的强化学习问题,构建基于DDPG算法的导弹规避训练系统;最后将基于DDPG算法的导弹规避训练系统自主学习到的逃逸机动策略与四种基于专家先验知识的经典逃逸机动进行对比,以验证基于DDPG算法的逃逸机动策略的有效性。

1 导弹飞机追逃模型

本文使用的导弹飞机追逃模型包括:飞机和导弹的质点模型、导弹的制导律模型、杀伤率模型以及导弹飞机相对运动模型等。

飞机规避导弹追逃模型的假设条件包括:(1)飞机和导弹都使用质点模型,考虑飞机和导弹的升阻特性和推力特性;(2)导弹采用比例导引制导律;(3)不考虑风的影响;(4)忽略侧滑角[9,22]。

1.1 飞机和导弹的质点模型

飞机和导弹的质点运动学模型为

$$\begin{cases}\dot{x}=V\cos\gamma\cos\chi\\ \dot{y}=V\cos\gamma\sin\chi\\ \dot{z}=-V\sin\gamma\end{cases}\tag{1}$$

式中:x,y,z分别为地轴系的三轴坐标,x轴指向正北,y轴指向正东,z轴竖直向下;V为飞行器飞行速度;$\dot{x},\dot{y},\dot{z}$为V在地轴系三个轴上的分量;γ为爬升角,表示速度和水平面夹角;χ为航迹方位角,表示飞行器飞行速度V在水平面的投影与x轴的夹角。

飞行器的质点动力学模型表示为[22]

$$\begin{cases}\dot{V}=g(n_{\mathrm{t}}-\sin\gamma)\\ \dot{\chi}=g\dfrac{n_{\mathrm{n}}\sin\mu}{V\cos\gamma}\\ \dot{\gamma}=\dfrac{g}{V}(n_{\mathrm{n}}\cos\mu-\cos\gamma)\end{cases}\tag{2}$$

式中:$\dot{V},\dot{\gamma},\dot{\chi}$分别为飞行器速度变化率、爬升角变化率以及航迹方位角变化率;$n_{\mathrm{t}},n_{\mathrm{n}},\mu$为飞行器的控制量,其中$n_{\mathrm{t}}$为沿速度方向的切向过载,控制飞行器的加减速;$n_{\mathrm{n}}$为沿飞行器升力方向的法向过载,控制升力方向上的运动;μ为航迹倾斜角;g为重力加速度。

$$\begin{cases}n_{\mathrm{t}}=\dfrac{T\cos\alpha-\overline{D}}{mg}\\ n_{\mathrm{n}}=\dfrac{T\sin\alpha+L}{mg}\end{cases}\tag{3}$$

式中:$L,\overline{D}$分别为升力和阻力;T为发动机推力;m为飞机质量;α为飞行器迎角。

在导弹飞机追逃模型中,受飞机升力、阻力和推力的限制,飞机切向过载$n_{\mathrm{tt}}\in[-2,1]$,法向过载$n_{\mathrm{nt}}\in[-4,8]$,航迹倾斜角$\mu_{\mathrm{t}}\in[-\pi,\pi]$。(下标t,m分别表示飞机(目标)和导弹。)

1.2　导弹飞机相对运动模型

导弹飞机相对运动示意图如图1所示,$[x_{\mathrm{m}},y_{\mathrm{m}},z_{\mathrm{m}}]^{\mathrm{T}}$,$[x_{\mathrm{t}},y_{\mathrm{t}},z_{\mathrm{t}}]^{\mathrm{T}}$分别表示导弹和飞机在地轴系的位置矢量,$\boldsymbol{D}$表示导弹相对飞机的位移,也称为瞄准线。

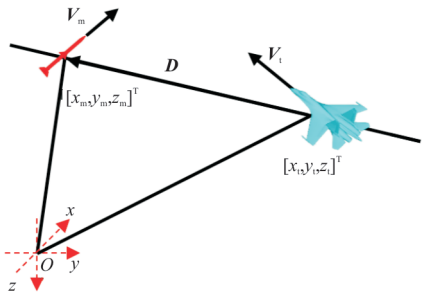


图1　导弹飞机相对运动模型

Fig. 1　Missile-aircraft engagement geometry model

$$\boldsymbol{D}=[x_{\mathrm{m}}-x_{\mathrm{t}},y_{\mathrm{m}}-y_{\mathrm{t}},z_{\mathrm{m}}-z_{\mathrm{t}}]^{\mathrm{T}}\tag{4}$$

导弹与飞机的距离:

$$D=|\boldsymbol{D}|\tag{5}$$

$\boldsymbol{V}_{\mathrm{m}},\boldsymbol{V}_{\mathrm{t}}$分别为导弹和飞机速度矢量,导弹相对飞机速度:

$$\Delta\boldsymbol{V}_{\mathrm{m}}=\boldsymbol{V}_{\mathrm{m}}-\boldsymbol{V}_{\mathrm{t}}\tag{6}$$

瞄准线变化率(导弹和飞机远离为正)为

$$\dot{D}=\frac{\Delta\boldsymbol{V}_{\mathrm{m}}\cdot\boldsymbol{D}}{D}\tag{7}$$

瞄准线角速度矢量:

$$\boldsymbol{\Omega}=\frac{\boldsymbol{D}\times\Delta\boldsymbol{V}_{\mathrm{m}}}{D^2}\tag{8}$$

瞄准线角速度大小:

$$\Omega=|\boldsymbol{\Omega}|\tag{9}$$

使用水平面内的飞机前置角和导弹进入角,如图2所示,其中$\boldsymbol{V}_{\mathrm{t}}'$和$\boldsymbol{V}_{\mathrm{m}}'$分别是飞机和导弹的速度在水平面的投影。

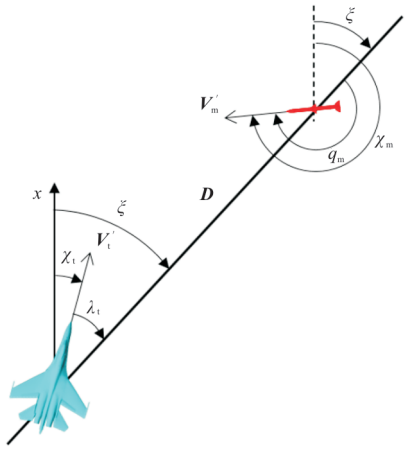


图2　飞机前置角和导弹进入角

Fig. 2　The aircraft bearing angle and the missile aspect angle

瞄准线方位角可表示为

$$\xi=\tan^{-1}\frac{y_{\mathrm{m}}-y_{\mathrm{t}}}{x_{\mathrm{m}}-x_{\mathrm{t}}}\tag{10}$$

水平面内飞机前置角(飞机速度与瞄准线之间的夹角):

$$\lambda_{\mathrm{t}}=\xi-\chi_{\mathrm{t}}\tag{11}$$

式中:χ_{t}为飞机航迹方位角。

水平面内导弹进入角(导弹速度方向与瞄准线之间的夹角):

$$q_{\mathrm{m}}=\xi-\chi_{\mathrm{m}}\tag{12}$$

式中:χ_{m}为导弹航迹方位角。

1.3　导弹导引律与杀伤率模型

空空导弹采用比例导引律,比例系数为

$$k = C\frac{|\dot{D}|}{V_m} \tag{13}$$

式中：V_m为导弹飞行速度大小；C为常数。

当瞄准线距离小于导弹杀伤半径或小于1.5倍距离变化率与时间步长的乘积时，判定导弹命中飞机。

$$(D < kr) \vee (D < 1.5|\dot{D}|\Delta t) \tag{14}$$

式中：kr为导弹杀伤半径；Δt为仿真时间步长；$\vee$为数学符号“或”。

本文忽略导弹和飞机的探测传感器以及电子对抗模型，只从运动学和动力学的角度考虑导弹失效条件，当且仅当导弹远离目标($\dot{D} > 0$)时，导弹失效。

2 DDPG算法

强化学习是智能体通过试错的机制和环境交互，目标是找到一个最优策略使得从环境中得到最大化的总奖励。强化学习可以被建模成一个马尔科夫过程(S, A, P, R)，其中S表示状态集合，A表示动作集合，P表示状态迁移模型，R表示奖励函数。在时间步长t内，智能体处于$s_t \in S$状态，根据策略π采取动作$a_t \in A$，收到奖励r_t。环境响应动作a_t，并向智能体呈现新的状态$s_{t+1} \in S$。时间步长t的总奖励为$R_t = \sum_{t'=t}^{\infty} \gamma^{t'-t} r_{t'}$，其中$\gamma \in [0, 1]$为折扣率。智能体的目的是学习到一个能最大化期望奖励的策略[23]。

策略π下状态s_t采取动作a_t的期望动作值函数：

$$Q^{\pi}(s_t, a_t) = E_{\pi}[R_t | s_t, a_t] \tag{15}$$

利用贝尔曼方程递归迭代更新估计动作值函数Q，直到找到最优策略。动作值函数使用贝尔曼方程估计[23]：

$$\begin{aligned} &Q^{\pi}(s_t, a_t) = \\ &E_{r_t, s_{t+1} \sim E}[r(s_t, a_t) + \gamma Q^{\pi}(s_{t+1}, a_{t+1})] \end{aligned} \tag{16}$$

DDPG算法是一种不依赖模型的基于actor-critic架构的深度强化学习算法，由策略网络和动作值网络构成。其中，确定性策略$\mu(s_t|\theta_\mu)$由参数为θ_μ的神经网络表示，动作值函数$Q(s_t, a_t|\theta_Q)$由参数为θ_Q的神经网络表示[17,24]。

critic网络的输出标签由贝尔曼方程估计得到，用y_t表示：

$$y_t = r(s_t, a_t) + \gamma Q[s_{t+1}, \mu(s_{t+1}|\theta_\mu)|\theta_Q] \tag{17}$$

critic网络的损失：

$$L(\theta_Q) = [Q(s_t, a_t|\theta_Q) - y_t]^2 \tag{18}$$

critic网络根据式(18)使用反向传播方法，对参数θ_Q进行优化。

actor网络使用策略梯度优化，策略梯度指预期收益函数J对策略函数参数θ_μ的梯度[17]

$$\begin{aligned} &\nabla_{\theta_\mu} J = E_{s \sim \rho^\mu}[\nabla_{\theta_\mu} Q(s_t, a_t|\theta_Q)|_{a_t = \mu(s_t|\theta_\mu)}] = \\ &E_{s \sim \rho^\mu}[\nabla_a Q(s_t, a_t|\theta_Q)|_{a_t = \mu(s_t)} \nabla_{\theta_\mu} \mu(s_t|\theta_\mu)] \end{aligned} \tag{19}$$

式中：ρ^μ为确定性策略的状态分布。

D. Silver等[25]证明了若$\nabla_{\theta_\mu}\mu(s_t|\theta_\mu)$和$\nabla_a Q(s_t, a_t|\theta_Q)$存在，则确定性策略梯度$\nabla_{\theta_\mu} J$存在。

DDPG借鉴深度Q网络(Deep Q Network)的经验池技术，把每一步的经验$e = (s_t, a_t, r_t, s_{t+1})$存储在经验池$D = \{e_1, e_2, \ldots e_m\}$中。由于在计算critic网络的损失函数时，$y_t$依赖于Q网络，而Q网络同时也在训练，会造成训练过程不稳定。因此，在actor和critic中分别建立目标网络对当前动作进行估计。目标网络延迟更新，训练更稳定，收敛性更好。

3 奖励设计

不同于监督学习可以使用标签，强化学习必须通过尝试去发现采取何种策略才能获取最大的奖励，因此稀疏奖励问题是深度强化学习应用于实际的核心问题。稀疏奖励在强化学习任务中广泛存在。智能体只有在完成任务时，才能获得奖励，中间过程无法获得奖励[26]。

本文增加人为设计的“密集”奖励，也被称为成型奖励。在智能体完成任务的过程中，通过成型奖励引导飞机成功规避导弹。

3.1 成型奖励设计

在智能体未得到最终结果之前，成型奖励可以评价智能体的策略。因此成型奖励的设计要准确，否则将导致策略函数收敛到局部最优。越复

杂的强化学习问题,影响奖励的因素越多,成型奖励设计的难度越大。

本文利用导弹飞机相对态势参数设计成型奖励(式(20)),引导飞机规避导弹,加快算法的收敛速度。

$$\begin{cases} r_d = C_1 \dfrac{D}{D_0} \\ r_{\dot{d}} = \dfrac{C_2 \dot{D}}{|\dot{D}|_{\max}} \\ r_\lambda = C_3 |\lambda_t| \\ r_q = C_4 |q_m| \\ r_{Ma} = C_5 Ma \\ r_h = C_6 h \\ r_s = r_d + r_{\dot{d}} + r_\lambda + r_q + r_{Ma} + r_h \end{cases} \tag{20}$$

式中:D_0为导弹与飞机初始距离;$|\dot{D}|_{\max}$为瞄准线变化率绝对值的最大值;$C_{i,i=1,2,\cdots,6}$分别为各分项奖励在总奖励中的权重系数;r_d为导弹与飞机距离奖励,$r_{\dot{d}}$为导弹与飞机距离变化率奖励,(r_d和$r_{\dot{d}}$描述了飞机导弹的距离态势);r_λ为飞机前置角奖励,r_q为导弹进入角奖励,(r_λ和r_q描述了飞机导弹的角度态势);r_{Ma}为飞机飞行马赫数奖励,其目的是防止飞机失速;r_h为飞机飞行高度奖励,其目的是防止飞机撞地。

3.2　稀疏奖励设计

稀疏奖励存在于绝大部分强化学习问题中,即任务完成后的奖励。飞机规避导弹问题的稀疏奖励是导弹飞机的交战结果,表示为

$$r_{\text{termin}} = \begin{cases} C_7 \gamma_t^{t_m - t} & (\text{missed}) \\ 0 & (\text{hit}) \end{cases} \tag{21}$$

式中:$\gamma_t \in [0,1]$为折扣系数;t_m为交战结果产生的时刻,即终局时刻;$t \in [0, t_m]$为当前时刻;missed表示飞机规避成功;hit则表示导弹成功命中飞机;C_7为成功规避奖励的权重系数。

结合式(20)~式(21),可得到总奖励表达式:

$$r_{\text{total}} = r_s + r_{\text{termin}} \tag{22}$$

式(20)和(21)中各奖励权重系数$C_{i,i=1,2,\cdots,7}$需要结合奖励参数的取值范围确定。

首先选取的是C_7的值,因为根据式(21),规避成功奖励为$C_7 \cdot 1$,被击中或坠毁的奖励为0。本文选取$C_7 = 40$,即成功逃逸后的奖励为40。根据$C_7 = 40$综合考虑式(20)各个奖励的取值范围,$C_{i,i=1,2,\cdots,6}$的值如表1所示。

表1　奖励系数值

Table 1　The values of reward coefficients

奖励参数范围	权重系数值	奖励范围
$\frac{D}{D_0} \in [0,1]$	$C_1 = 15$	$r_d \in [0,15]$
$\frac{\dot{D}}{\lvert\dot{D}\rvert_{\max}} \in [-1,1]$	$C_2 = 30$	$r_{\dot{d}} \in [-30,30]$
$\lvert\lambda_t\rvert \in [0,180]$	$C_3 = 0.05$	$r_\lambda \in [0,20]$
$\lvert q_m\rvert \in [0,180]$	$C_4 = -0.05$	$r_q \in [-20,0]$
$Ma \in [0,2.0]$	$C_5 = 5$	$r_{Ma} \in [0,10]$
$h \in [0,10\,000]$	$C_6 = 0.001$	$r_h \in [0,10]$

4　基于DDPG的飞机规避导弹训练系统框架

把飞机规避导弹问题建模为一个强化学习问题,飞机与导弹在时刻t的运动参数和相对态势为强化学习的状态s_t,飞机t时刻的控制指令为强化学习的动作a_t,导弹飞机的追逃模型是强化学习的环境。

基于DDPG强化学习方法、导弹飞机追逃模型及飞机规避导弹的奖励设计,可建立飞机规避导弹训练系统,如图3所示。

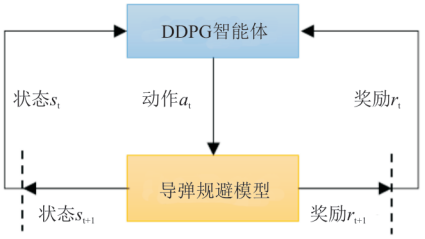


图3　基于DDPG的飞机规避导弹训练系统

Fig. 3　The missile evasion training system based on the DDPG

基于DDPG的飞机规避导弹训练系统的状态s_t共8个,如表2所示。动作a_t共3个,如表3所示。

表2和表3中的状态s_t和动作a_t分别为actor网络的输入和输出。输入和输出的量都在-1到1

之间，输入输出根据各自取值范围进行归一化和反归一化。

表 2 飞机规避导弹训练系统的状态

Table 2 The states of the missile evasion training system

状态符号	状态描述
D_0	飞机导弹初始距离
D	瞄准线距离
$\dot{D}$	瞄准线变化率
λ_t	飞机前置角
q_m	导弹进入角
$\lvert\Omega\rvert$	瞄准线旋转角速率
Ma	飞机飞行马赫数
h	飞机飞行高度

表 3 飞机规避导弹训练系统的动作

Table 3 The actions of the missile evasion training system

动作符号	动作描述
n_{tt}	飞机切向过载指令
n_{nt}	飞机法向过载指令
μ_t	飞机航迹倾斜角指令

基于DDPG的智能体只依赖飞机规避导弹环境产生并存储在经验池中的经验数据和式(22)的奖励设计，在没有其他先验知识的情况下，通过训练找到行之有效的逃逸机动策略。

5 仿真过程与结果

5.1 初始场景设置

空空导弹攻击区是指空空导弹发射时刻能够命中目标的空间区域。导弹攻击区与许多因素有关，包括导弹和飞机的初始速度和高度、导弹离轴发射角、目标进入角、导弹制导律和目标机动方式等。它不仅是衡量空空导弹攻击能力的指标，也是衡量目标飞机逃逸机动策略有效性的指标。在其他影响因素相同的前提下，飞机逃逸机动对应的导弹攻击区越小表明飞机逃逸机动越有效。

考虑到超视距空战场景中，导弹迎面发射对目标飞机的威胁最大，因此本文将攻击区的范围限制在飞机前置角－30°～30°的范围内，本文使用如图4所示的攻击区作为逃逸机动策略的评价标准。

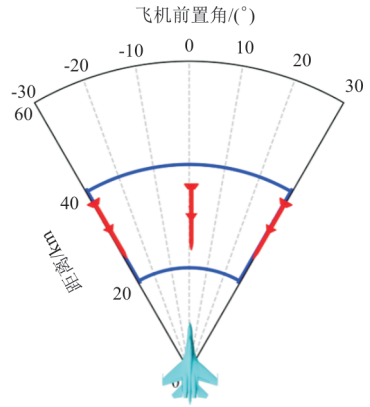


图4 飞机前置角－30°～30°攻击区

Fig. 4 The attack zone of aircraft bearing angle ranging from －30° to 30°

智能体训练的初始场景配置如下：飞机位置不变，始终在原点，航向正北。导弹初始位置在以飞机为圆心，半径20～40 km，飞机前置角－30°～30°的闭合范围内(图4闭合线部分)，导弹的航向始终指向飞机。考虑导弹从载机发射，导弹初始高度和马赫数取决于载机的高度和马赫数。因此本文设置飞机导弹的初始高度都为8 000 m，初始马赫数都为0.9。导弹发射后会急剧加速，导弹马赫数随飞行时间变化如图5所示，最大马赫数大于5.0。

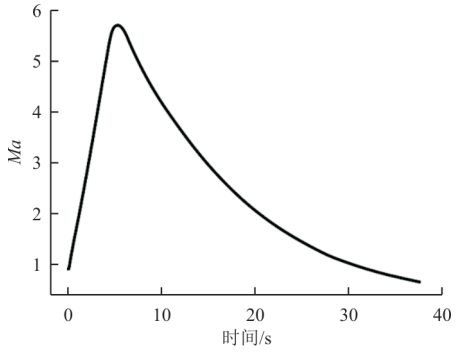


图5 导弹飞行马赫数随时间变化

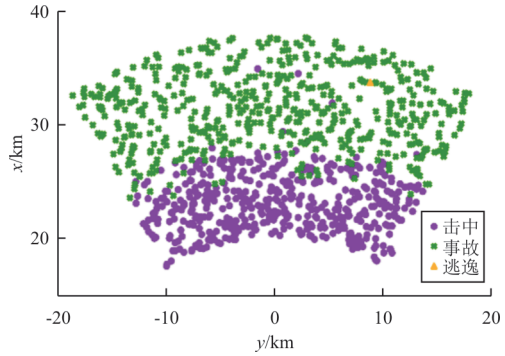
Fig. 5 The missile Mach number changes with time

5.2 训练过程与结果

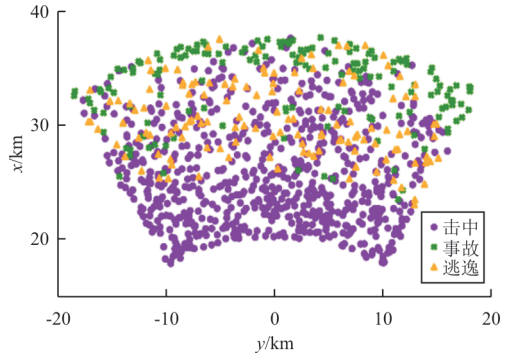
总共训练约70万次，仿真共生成2.3亿组经验($e=(s_t, a_t, r_t, s_{t+1})$)数据。

训练过程与结果如图6所示，x形符号表示飞机以非正常飞行状态(失速或撞地)结束仿真，圆

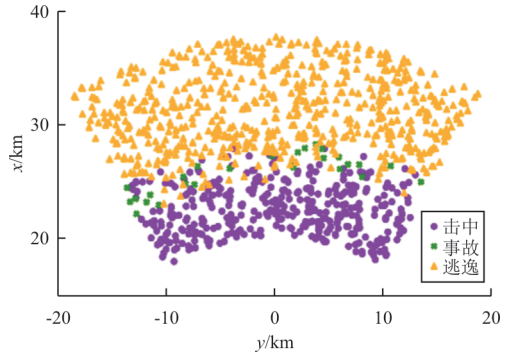
形符号表示飞机被导弹击中,三角符号表示飞机成功规避导弹。



(a) 前 1 000 代的训练结果



(b) 2 000~3 000 代的训练结果



(c) 最后 1 000 代的训练结果

图 6　训练过程与结果

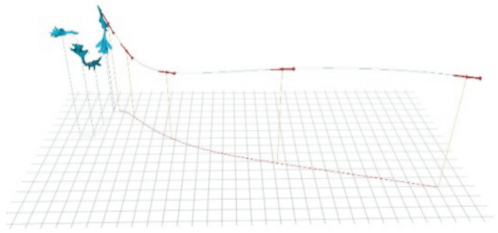
Fig. 6　Training process and results

从图 6(a)可以看出:在前 1 000 代训练中,飞机失速或撞地占 52.6%,被导弹击中占 47.3%,飞机只成功规避一次导弹。此时智能体尚未学会控制飞机正常飞行,更无法规避导弹。

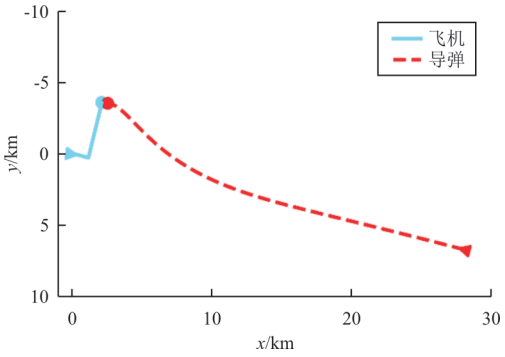
第 2 000 代~3 000 代的训练结果如图 6(b)所示,可以看出:飞机失速或撞地占 15.5%,被导弹击中次数占 70.9%,飞机成功规避导弹占 13.6%。此时智能体已能逐渐控制飞机飞行,然而还未能有效规避导弹。

最后 1 000 代的训练结果如图 6(c)所示,可以看出:飞机失速或撞地只占总数的 2.7%,被导弹击中占 37.0%,飞机成功规避导弹占总数的 60.3%。此时智能体已自主学习到一种飞机规避导弹策略,能在约 25 km 外规避导弹。

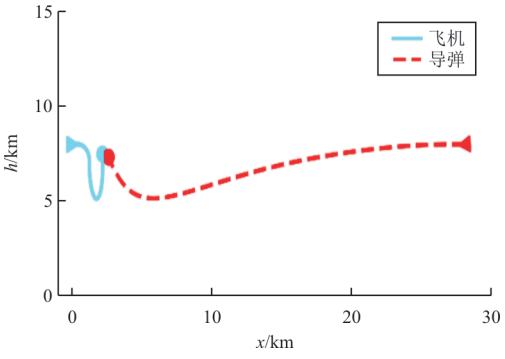
智能体学习到的逃逸机动策略如图 7 所示,实线为飞机,虚线为导弹。图 7(a)是逃逸机动的三维轨迹图,图 7(b)和图 7(c)是以地轴系 x 坐标为横坐标,地轴系 y 坐标和高度 h 分别为纵坐标的飞机导弹飞行轨迹图。三角形表示起点,实心圆表示终点。图 7(b)中 y 轴正方向向下。



(a) 逃逸机动三维轨迹图



(b) 逃逸机动轨迹在水平面投影



(c) 逃逸机动轨迹在 x-h 平面投影

图 7　基于 DDPG 算法的逃逸机动策略

Fig. 7　The evasive maneuver strategy based on the DDPG algorithm

从图 7 可以看出：智能体实现的逃逸机动策略为导弹发射后，飞机急剧转弯，尽快把导弹置于尾后，转弯的同时降低高度直至 5 000 m 左右，最后拉起飞机成功规避导弹。

5.3 典型逃逸机动策略介绍

参考文献[4-5,7,27]对飞机规避导弹问题的研究，总结出四种典型的依赖导弹规避先验知识的逃逸机动策略。

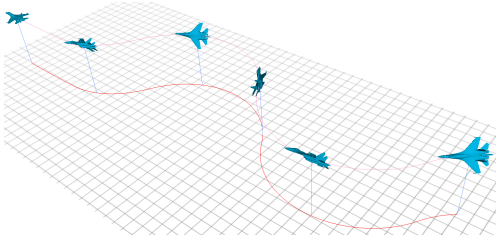
（1）定直平飞：逃逸飞机保持初始高度、速度和航向飞行。

（2）蛇形机动：逃逸飞机航迹方位角 χ 在一定幅值范围内连续周期性变化。

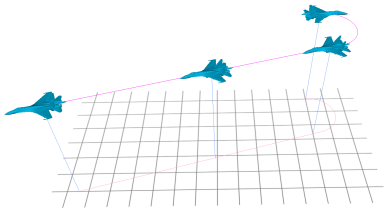
（3）水平置尾机动：逃逸飞机以最大稳定盘旋角速度转弯至置尾（飞机与导弹航向偏差小于 5°），然后以加力状态平飞逃逸。

（4）置尾下降机动：逃逸飞机以大于 90°滚转角边转弯边下降直至飞机与导弹航向偏差小于 5°，后改出下降在低空以加力状态平飞逃逸。

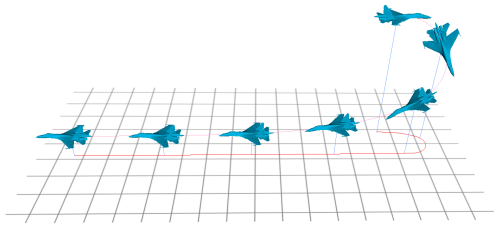
典型逃逸机动策略如图 8 所示。



（a）蛇形机动



（b）水平置尾机动



（c）置尾下降机动

图 8 典型逃逸机动策略

Fig. 8 The classic evasive maneuver strategies

与智能体训练导弹规避的初始条件相同，各个典型逃逸机动的攻击区，如图 9 所示。

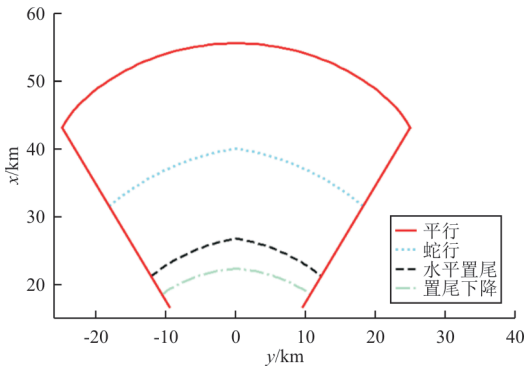


图 9 典型机动策略下的攻击区

Fig. 9 The attack zones of classic maneuver strategies

结合图 6(c)和图 9 得到典型机动策略和智能体自主学习的逃逸机动策略的攻击区对比图，如图 10 所示。

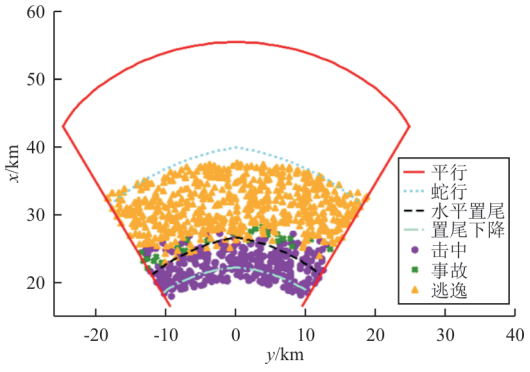


图 10 所有逃逸机动策略攻击区对比图

Fig. 10 The attack zones of all evasive maneuver strategies

从图 10 可以看出：所有逃逸机动策略的攻击区从大到小依次为：平飞机动＞蛇形机动＞水平置尾机动≈智能体实现的逃逸机动＞置尾下降机动。

综上所述，利用深度确定性策略算法实现的逃逸机动，在没有任何飞机规避导弹先验知识的情况下，攻击区优于蛇形机动，与水平置尾机动持平，稍劣于置尾下降机动。

6 结 论

（1）本文所构建的基于 DDPG 算法的导弹规避训练系统表明，智能体在不依赖导弹规避先验知识、仅凭借仿真数据和奖励的情况下，最终能够

自主学习到一种有效的逃逸机动策略。

(2) 通过与四种典型逃逸机动策略的攻击区相比,智能体逃逸机动攻击区仅次于置尾下降攻击区,但智能体实现的逃逸机动策略对导弹规避的先验知识需求最少。

参 考 文 献

[1] 樊会涛, 崔颢, 天光. 空空导弹70年发展综述[J]. 航空兵器, 2016 (1): 43171.
FAN Huitao, CUI Hao, TIAN Guang. A review on the 70-year development of air-to-air missiles[J]. Aero Weaponry, 2016 (1): 43171. (in Chinese)

[2] SHINAR J, GUELMAN M, SILBERMAN G, et al. On optimal missile avoidance-a comparison between optimal control and differential game solutions[C]// Proceedings ICCON IEEE International Conference on Control and Applications. Jerusalem: IEEE, 1989: 453-459.

[3] BURGIN G, WILLIAMS W, SIDOR L. The adaptive maneuvering logic program in support of the pilot's associate program-a heuristic approach to missile evasion[C]// 24th Aerospace Sciences Meeting. Reno: AIAA, 1986: 423.

[4] MANDT G, NEIGHBOUR T. Air-to-air missile avoidance[C]// Guidance and Control Conference. San Diego: AIAA, 1982: 1516.

[5] 邵彦昊, 朱荣刚, 贺建良, 等. 中远程空空雷达导弹的新机动规避方式的探索[J]. 弹箭与制导学报, 2020, 40(4): 75-78,84.
SHAO Yanhao, ZHU Ronggang, HE Jianliang, et al. Exploration of a new evasive maneuver mode for medium and long range air-to-air radar missile[J]. Journal of Projectiles, Rockets, Missiles and Guidance, 2020, 40(4): 75-78,84. (in Chinese)

[6] LAPUMA A, MARLIN C. Pilot's associate-a synergistic system reaches maturity[C]// 9th Computing in Aerospace Conference. San Diego: AIAA, 1993: 4665.

[7] 王斯财, 南英, 刘经纬. 导弹迎击时飞机的最佳逃逸策略研究[J]. 航空兵器, 2009(4): 28-32.
WANG Sicai, NAN Ying, LIU Jingwei. Optimal escape strategy of fighter against oncoming missiles[J]. Aero Weaponry, 2009(4): 28-32. (in Chinese)

[8] IMADO F. Some aspects of a realistic three-dimensional pursuit-evasion game[J]. Journal of Guidance, Control, and Dynamics, 1993, 16(2): 289-293.

[9] RAIVIO T. Capture set computation of an optimally guided missile[J]. Journal of Guidance, Control, and Dynamics, 2001, 24(6): 1167-1175.

[10] IMADO F, KURODA T. Family of local solutions in a missile-aircraft differential game[J]. Journal of Guidance, Control, and Dynamics, 2011, 34(2): 583-591.

[11] IMADO F. Some practical approaches to pursuit-evasion dynamic games[J]. Cybernetics and Systems Analysis, 2002, 38(2): 276-291.

[12] IMADO F, KURODA T. Engagement tactics for two missiles against an optimally maneuvering aircraft[J]. Journal of Guidance, Control, and Dynamics, 2011, 34(2): 574-582.

[13] ONG S Y, PIERSON B L. Optimal evasive aircraft maneuvers against a surface-to-air missile[C]// Proceedings The First IEEE Regional Conference on Aerospace Control Systems. Westlake Village: IEEE, 1993: 475-482.

[14] SINGH L. Autonomous missile avoidance using nonlinear model predictive control[C]// AIAA Guidance, Navigation, and Control Conference and Exhibit. Providence: AIAA, 2004: 4910.

[15] KARELAHTJ J, VIRTANEN K, RAIVIO T. Near-optimal missile avoidance trajectories via receding horizon control[J]. Journal of Guidance, Control, and Dynamics, 2007, 30(5): 1287-1298.

[16] MOUSAVI S, SCHUKAT M, HOWLEY E. Deep reinforcement learning: an overview[C]// Proceedings of SAI Intelligent Systems Conference. [S. l.]: SAI, 2018: 319-324.

[17] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning: United States, 20170024643 [P/OL]. [2021-02-08]. https://www.freepatentsonline.com/y2017/0024643.html..

[18] WANG M, WANG L, YUE T. An application of continuous deep reinforcement learning approach to pursuit-evasion differential game[C]// 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). Chengdu: IEEE, 2019: 1150-1156.

[19] YOU S, DIAO M, GAO L, et al. Target tracking strategy using deep deterministic policy gradient[J]. Applied Soft Computing, 2020, 95: 106490.

[20] CIMURS R, LEE J H, SUH I H. Goal-oriented obstacle avoidance with deep reinforcement learning in continuous action space[J]. Electronics, 2020, 9(3): 411.

[21] 范鑫磊, 李栋, 张尉, 等. 基于深度强化学习的导弹规避决策训练研究[J]. 电光与控制, 2021, 28(1):81-85.
FAN Xinlei, LI Dong, ZHANG Wei, et al. Missile evasion decision training based on deep reinforcement learning[J]. Electronics Optics & Control, 2021, 28(1): 81-85. (in Chinese)

[22] SHINAR J, GAZIT R. Optimal "no-escape" firing envelopes of guided missiles[C]// 7th Computational Fluid Dynamics Conference. Cincinnati: AIAA, 1985: 1960.

[23] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. 2nd ed. Cambridge: MIT Press, 2018.

[24] 黄旭, 柳嘉润, 贾晨辉, 等. 深度确定性策略梯度算法用于无人飞行器控制[J]. 航空学报, 2021, 42(X): 524688.
HUANG Xu, LIU Jiarun, JIA Chenhui, et al. Deep deterministic policy gradient for UAV control[J]. Acta Aeronautica et Astronautica Sinica, 2021, 42(X): 524688. (in Chinese)

[25] SILVER D, LEVER G, HEESS N, et al. Deterministic policy gradient algorithms[C]// Proceedings of the 31st International Conference on Machine Learning. Beijing: PMLR, 2014: 387-395.

[26] 杨惟轶, 白辰甲, 蔡超, 等. 深度强化学习中稀疏奖励问题研究综述[J]. 计算机科学, 2020, 47(3): 182-191.
YANG Weiyi, BAI Chenjia, CAI Chao, et al. Survey on sparse reward in deep reinforcement learning[J]. Computer Science, 2020, 47(3): 182-191. (in Chinese)

[27] YANG Z, ZHOU D, PIAO H, et al. Evasive maneuver strategy for UCAV in beyond-visual-range air combat based on hierarchical multi-objective evolutionary algorithm[J]. IEEE Access, 2020, 8: 46605-46623.

作者简介:

宋宏川(1989—),男,博士研究生。主要研究方向:飞行动力学与控制、飞行仿真、智能空战。

詹　浩(1972—),男,博士,教授。主要研究方向:新概念飞行器总体设计、高效多目标优化算法及其应用、飞行器复杂状态下的飞行动力学与控制、非定常气动力数值仿真技术。

夏　露(1977—),女,博士,副教授。主要研究方向:飞行器概念设计、气动隐身设计。

李向阳(1991—),男,硕士,工程师。主要研究方向:深度强化学习、飞行动力学与控制、智能空战。

刘　艳(1981—),女,博士,副教授。主要研究方向:飞行动力学与控制。

(编辑:马文静)